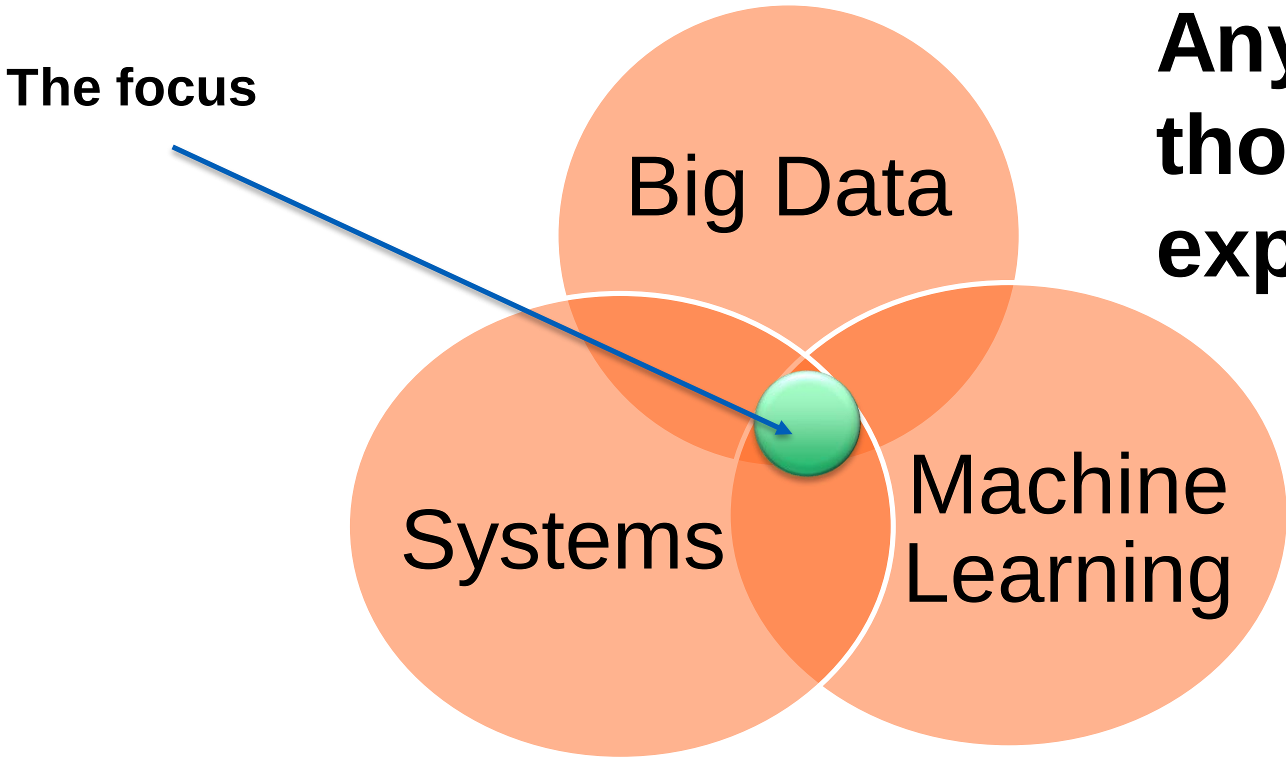
Aalto University
School of Science

Machine Learning with Edge Systems

Hong-Linh Truong
Department of Computer Science
linh.truong@aalto.fi, *https://rdsea.github.io*

Our focus in this course

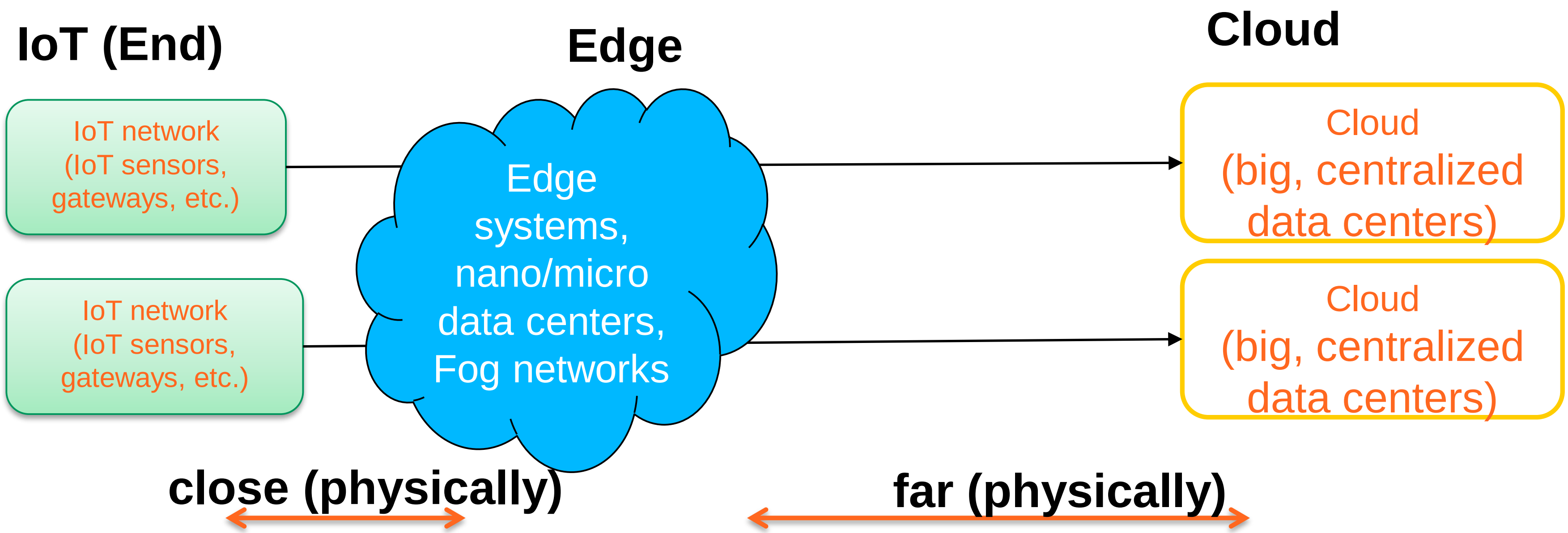


Any idea, thought, expectation?

Content

- **Edge computing**
- **Why would ML in the edge be our focus?**
- **Some open areas**
  - MLOps for edge systems
  - Transfer learning
  - Federated learning
  - Elastic serving/inferencing

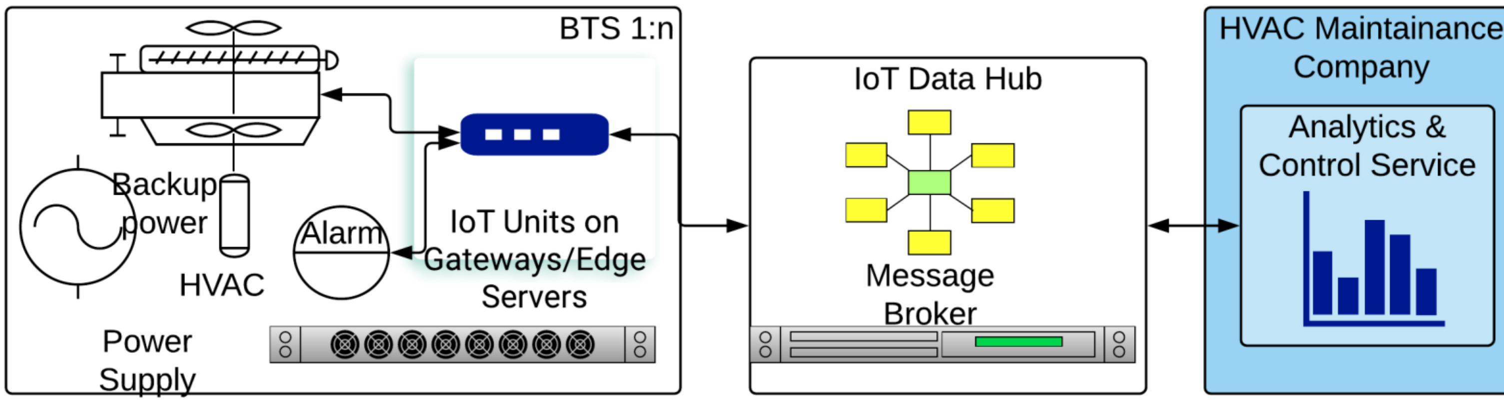
IoT-Edge-Cloud



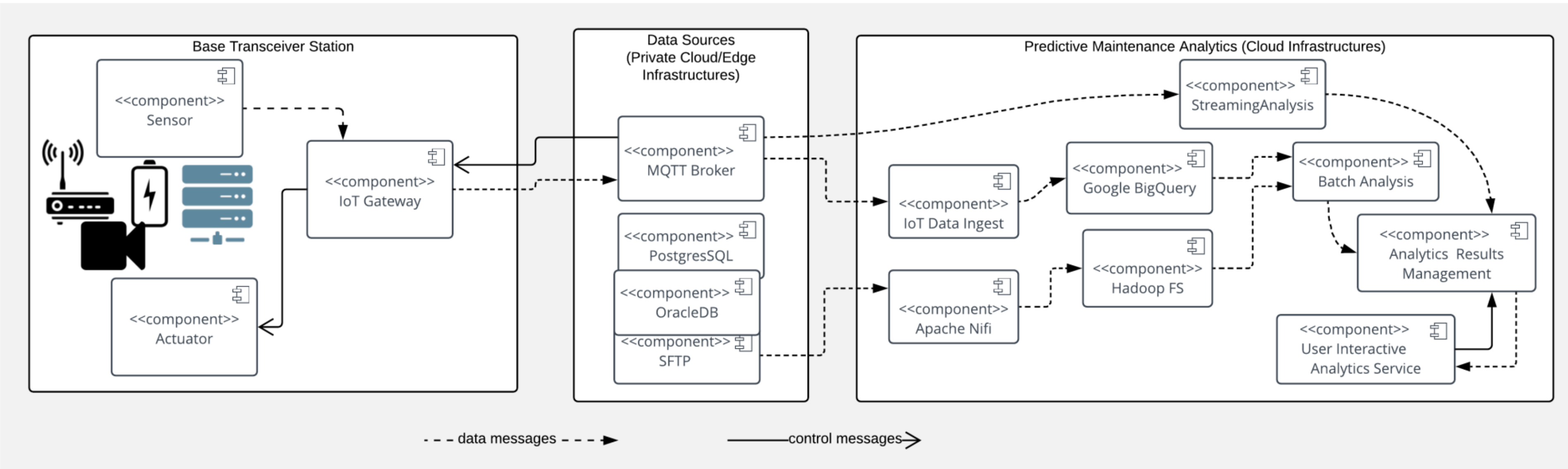
Edge computing

- **Edge computing paradigm focuses on distributed computing at the edge**
  - “Edge” is an abstraction
    - *Distributed large number of low-end devices as well as very limited high-end devices*
  - Common technologies like in the cloud and specific ones
    - *E.g., virtualized machines, message brokers, storage, Web services*
- **Computation/Analytics at the edge**
  - Where data is generated, close to the data sources
    - *Next to IoT devices and sensing equipment, E.g., in the shopping centrc, in the car*
  - Near real-time processing is needed

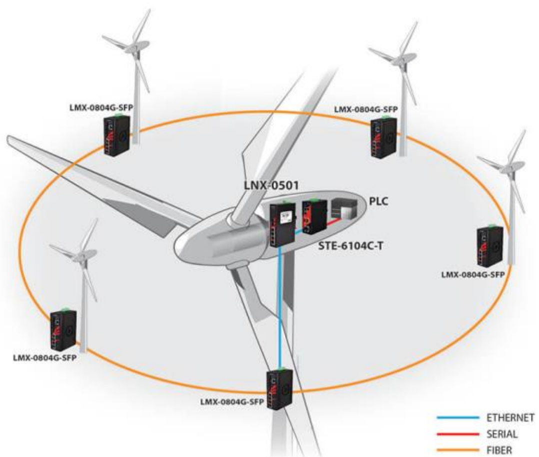
Predictive maintenance

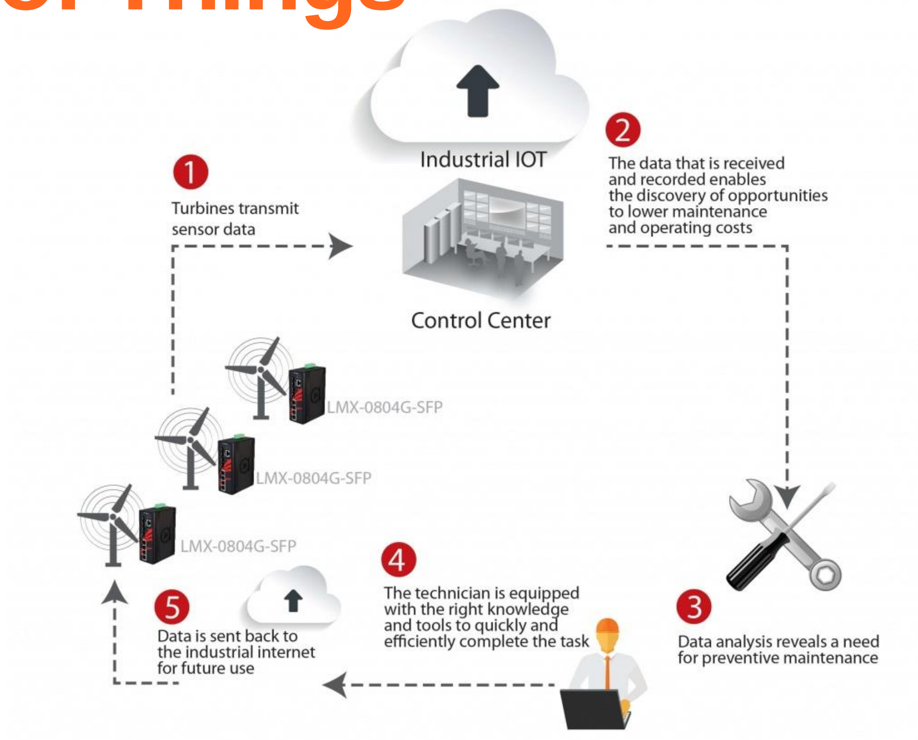


Predictive maintenance



Industrial Internet of Things





Figures source: http://www.windpowerengineering.com/design/electrical/controls/wind-farm-networks/talking-turbines-internet-things/

Video analytics at the edge

Use Case 3: Video Analytics

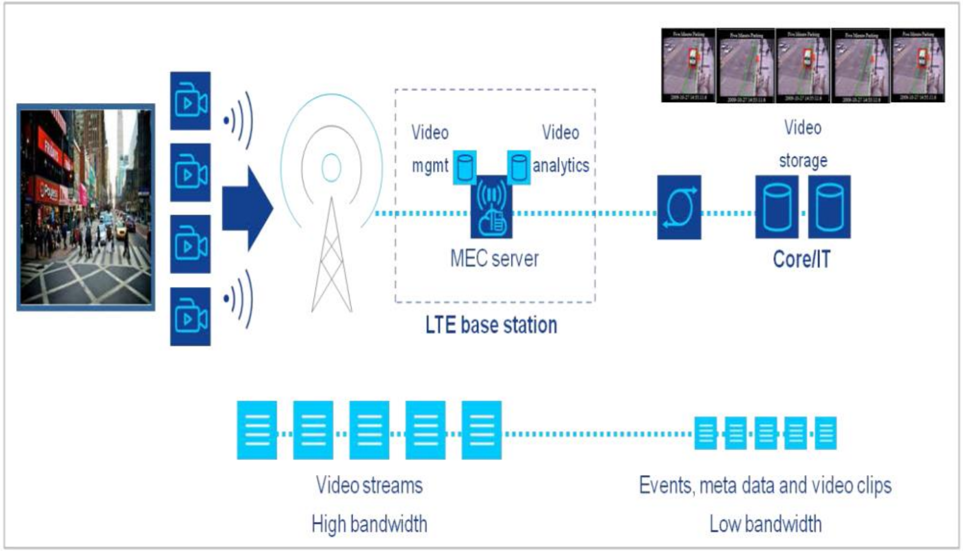


Figure 4: Example of video analytics

Figure source:
https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-_introductory_technical_white_paper_v1%2018-09-14.pdf

Why do we have to support ML at the edge?

Why do we have to support machine learning/big data in the edge

- **Close to data sources → "data locality" benefits**
  - Security & privacy
  - Performance
  - Customization
- **Many applications (AI is specific application anyway)**
  - Inferencing/classification in mobile devices
  - Realtime ML (autonomous cars, speech recognition, fraud detection)
  - Manufacturing (Industrial Internet of Things)
    - *Anomaly detection*

What do we need to consider when supporting ML in the edge?

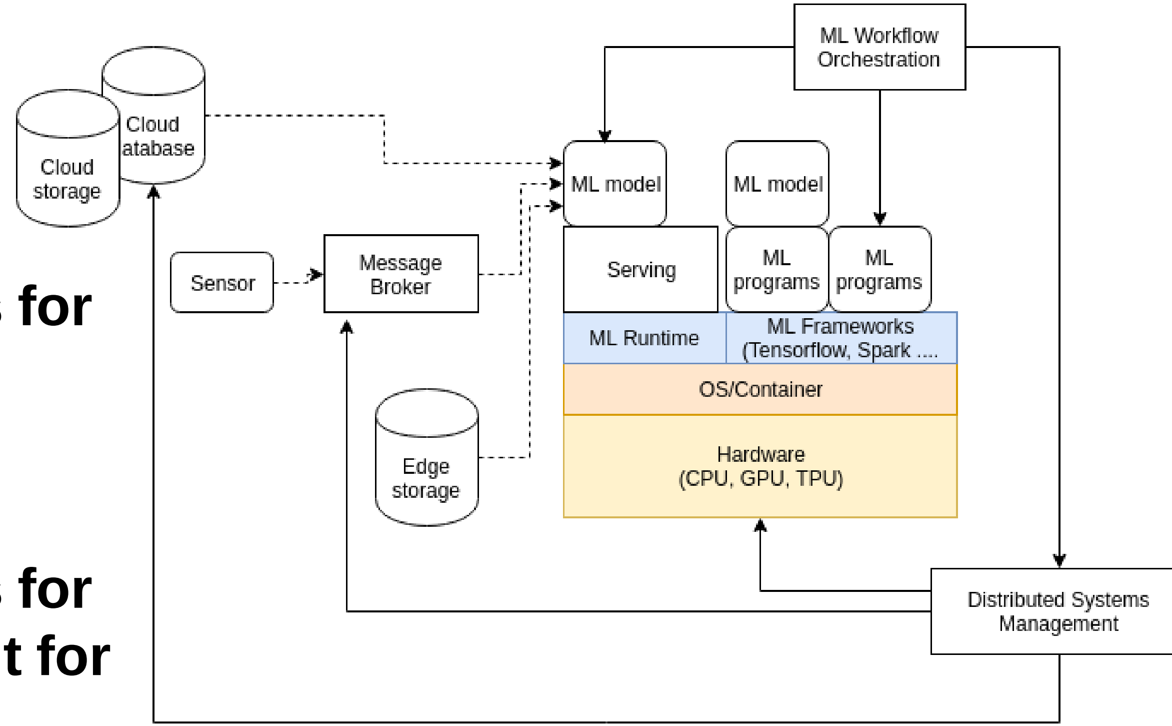
- **Network problems**
  - Low latency, low-bandwidth, unreliable connectivity
- **Computation capabilities**
  - Constrained power, a lot of specific chips and accelerators
  - Limited memory
- **Storage is not enough for big data**
- **Data**
  - Opportunistic data, unlabeled data, time series data
  - Streaming data

Imagine you work on ML in the edge systems (and you know ML in clouds already)

Pervasive embedded Edge devices

- Raspberry PI4
- Google Coral
- Jetson Nano
- Xilinx
- A huge number of MCUs (MicroController Units)

Software systems for ML in the edge



- What are key features for ML runtime and programming frameworks?
- What are key features for resource management for running ML?

Suitable ML and Runtime for the edge: Key requirements

- **Energy consumption**
- **Resource constraints**
  - less computation capabilities
- **Latency and uncertainty**
- **Interfaces with different networks capabilities**
- **Support accelerators**
  - E.g., FPGA, AI Accelerators (e.g. Intel® Movidius Myriad X VPU)
- **Trade-offs between generic versus specific features**

Examples of ML frameworks and Runtime for the edge

- **TF-lite**
  - https://www.tensorflow.org/lite
- **https://microsoft.github.io/ELL/**
- **https://github.com/Microsoft/EdgeML**
- **uTensor: https://github.com/uTensor/uTensor**
- **Androi NN**
- **CoreML 3**
- **PyTorch mobile**
- **Snapdragon Neural Processing Engine SDK**
  - https://developer.qualcomm.com/docs/snpe/overview.html

Changes in MLOps

- **MLOps (ML DevOps)**
  - DevOps principles for ML
  - In ML engineering processes: key artefacts are ML models, data and runtime libs
  - New areas, still a lot of ongoing research work
- **Changes in ML with edge systems**
  - DevOps and DataOps centered around models and data
  - Optimization and training activities
  - Tests and benchmarks
  - Monitoring

Example of Google PL

https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Is it the same in the edge?

What would be MLOps for ML in the edge?

MLOps in edge systems

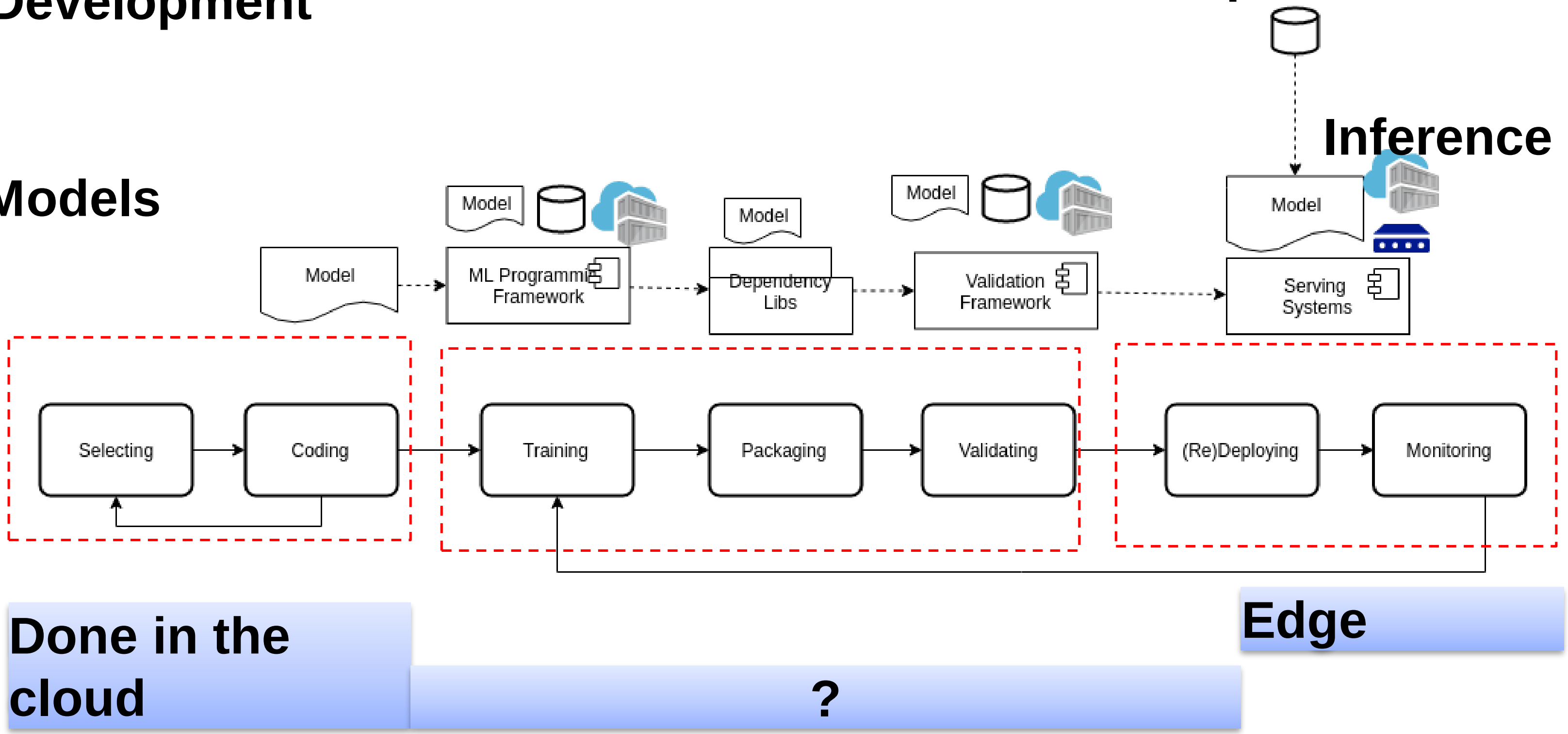
Development

Operations

Artefacts: Models framework

Phases Activities

Where?



Training in cloud and inference in the edge

https://blogs.gartner.com/paul-debeasi/files/2019/01/Train-versus-Inference.png

Can you guess some issues that you need to deal with in the MLOps for the edge?

Examples

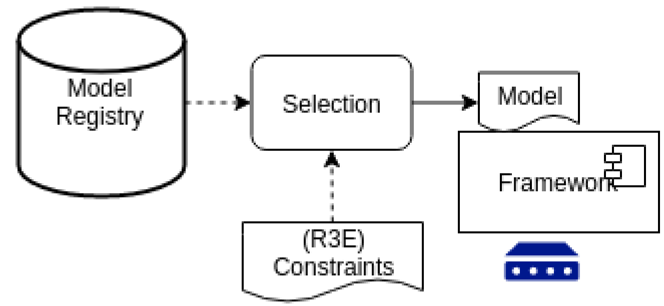
https://developer.qualcomm.com/docs/snpe/overview.html

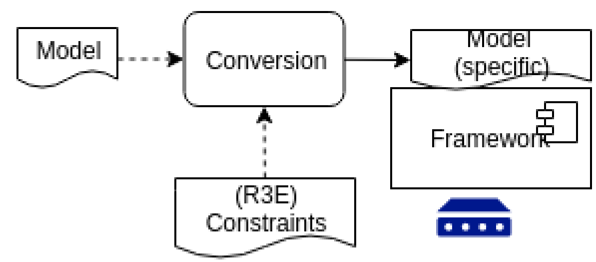
Selected problems: transfer learning

- **Transfer learning**
  - Repurpose a model trained for a task for another task
  - Basically it is an optimization of an existing model for a new task
- **Transfer learning for the edge**
  - Convert typical models to edge models
  - Need model selection, reuse and model retraining
  - Combine with other optimization techniques

Selected problems: model selection and conversion

- **Model management and selection**
  - Precision and time tradeoffs with computational requirements
  - Work with accelerators
- **Conversion**
  - A model can be supported by different frameworks
- **How will these issues affect Robustness and Reliability?**



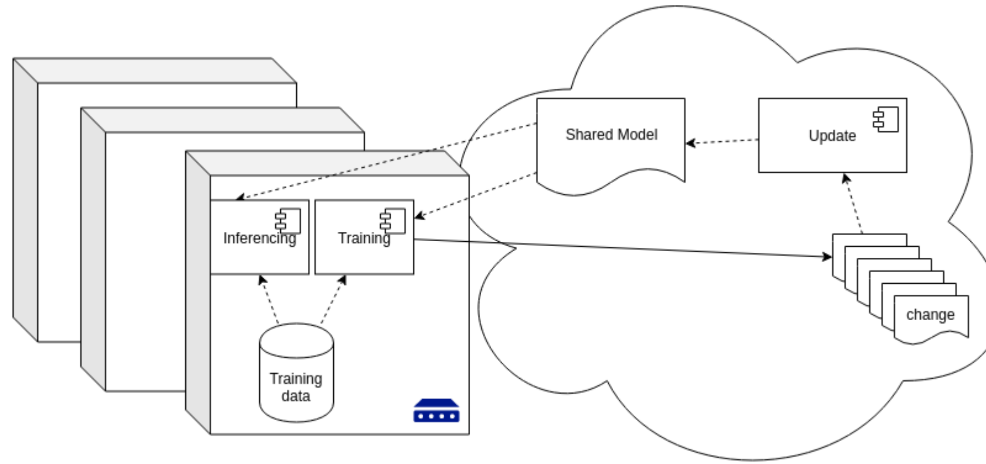


Selected problems: model optimization

- **Pruning**
  - Prune graphs for training, remove features in ML models which are not significant
- **Quantization**
  - Reduce precision representation, storage, bandwidth
- **Conditional computation/Regularization**
  - Activate certain units of the model
- **How will these issues affect Robustness, Reliability and Elasticity?**

Selected problems: federated training with edges

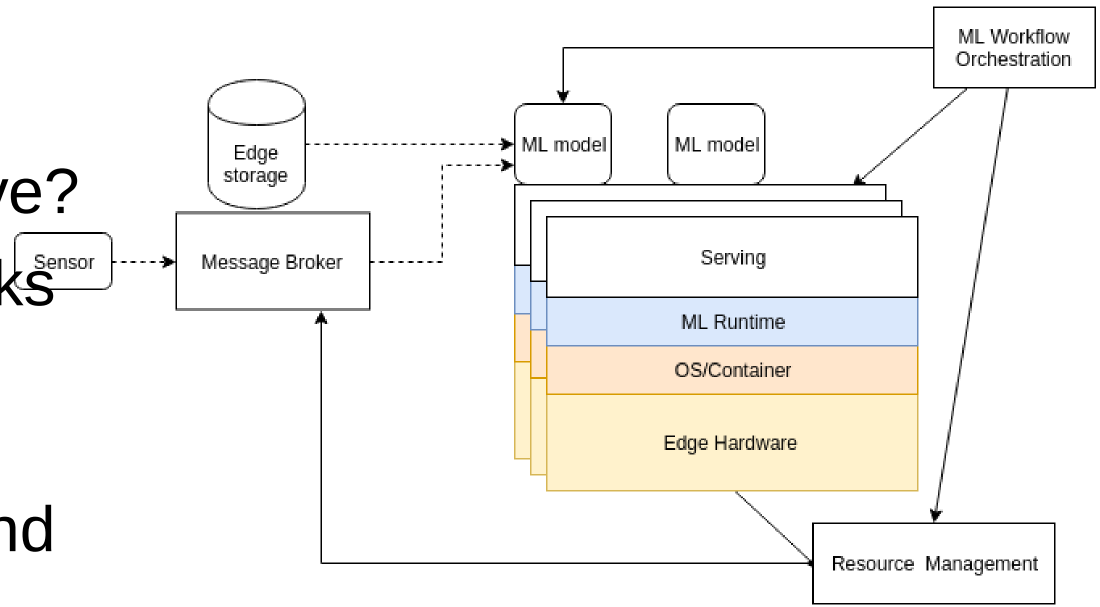
Machine learning is decentralized with a distributed set of devices holding data and carrying out (sub) training/inferencing



- **What about Reliability and Resilience?**
  - Consensus in updates, secured aggregation protocols, dynamicity and elasticity

Selected problems: ML Serving

- **ML Serving (and R3E)**
  - Which types of dynamic service models we could have?
  - How to distribute tasks in model serving?
  - How to partition ML tasks in both edge and cloud?



Study log

- **No study log but read papers to start working on ML for edge systems**
- **You can pickup some points mentioned as the topic for your individual project**
  - Or incorporate some ideas into your individual project
- **We expect ML with edge systems will be the main focuses soon in our advanced software systems course!**
  - Good areas for master theses/research projects.

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io